Large Language Models Bias Issues Solving Through SDRT

Nagesh Somayajula[1], Chinmay Somayajula [2]
[1]B.S. and M.S. in Science and Mathematics from Andhra University, India
[2] Independent researcher

ABSTRACT Since the start of transformer development and recent advancements in large language models (LLMs), the whole world has been taken by storm. However, multiple LLM models, such as GPT-3, GPT-4, and all open-source LLM models, come with their own set of challenges.

The development of Natural Language Processing (NLP) utilizing transformers commenced in 2017, initiated by Google and Facebook. Since then, substantial language models have emerged as formidable tools in the domains of both natural language and artificial intelligence research. These models possess the capability to learn and predict, enabling them to generate coherent and contextually relevant text for a diverse array of applications.

Additionally, large language models have made a significant impact on various industries, including healthcare, finance, customer service, and content generation. They have the potential to automate tasks, improve language understanding, and enhance user experiences when deployed effectively. However, along with these benefits, there are also major risks and challenges associated with these models, including pre-training and fine-tuning. To address these challenges, we are proposing SDRT (Segmented Discourse Representation Theory) and making the models more conversational to overcome some of the toughest obstacles.

Introduction

Furthermore, as we increasingly adopt Large Language Models (LLMs) in our daily lives, work, and enterprise automation, we encounter various challenges associated with these models. These challenges encompass computational requirements, ethical considerations, bias mitigation, and interpretability. They underscore the imperative of responsible development and deployment of these models to ensure fair and unbiased outcomes. To address these challenges, we propose an architectural change in existing transformer models, incorporating Structured Dialogic Response

Theory (SDRT) components on top of attention models.

Why is it crucial to implement this new architectural change? It is because large language models represent a significant advancement in natural language processing, offering unprecedented capabilities in text understanding and generation. With ongoing research and refinement, these models have the potential to revolutionize numerous fields, shaping the future of human-computer interaction and communication. However, they also have the potential to create issues such as fake news and bias, thereby underscoring the need for proactive solutions.

Some of the top challenges associated with large language models are:

Ethical Concerns: Large language models can generate highly convincing and coherent text, which raises concerns regarding the potential for misuse. There are worries about the dissemination of misinformation, the spreading of biased or harmful content, and the potential for deep fakes or impersonation.

Bias in Training Data: Language models learn from vast amounts of data available on the internet, which can contain biases present in the data sources. This can result in the models perpetuating or amplifying biases related to gender, race, or other sensitive topics.

Lack of Contextual Understanding: While large language models excel at generating text, they often lack true comprehension or contextual understanding. They rely heavily on patterns in the data they were trained on and may produce plausible-sounding but factually incorrect or nonsensical responses.

Data Privacy and Security: Language models may inadvertently memorize or reproduce sensitive information present in the training data, posing risks to data privacy and security. There are concerns about the potential exposure of personal or confidential information through generated text.

Explainability and Interpretability: Large language models are often considered black boxes, making it difficult to understand their decision-making processes or provide explanations for their outputs. This lack of transparency can be problematic in critical domains where accountability and interpretability are crucial.

- We propose addressing some of these challenges using SDRT, which stands for Segmented Discourse Representation Theory.
- In this proposal, we aim to integrate SDRT into existing transformer models by incorporating SDRT with attention logic in both the decoders and encoders. This modification is intended for various tasks, including translation, text-to-text generation, question-answering, and chatbots. By doing so, we aim to reduce bias in the data and promote more meaningful conversations, ensuring that the right audience receives the right information.
- Our demonstration shows that our approach remains robust even when dealing with missing features during inference, resulting in high-quality output by mitigating data bias and fostering more conversational representations.
- Additionally, we illustrate how each response generated by the machine becomes more transparent and explainable.

Background:

In the current architecture, when faced with a long or complex sentence, the system focuses solely on tokenizing and sending it to the decoder for text generation or understanding. It lacks an overall understanding of the sentence's purpose. To address this, the emphasis should shift to comprehending the semantics of each word in relation to the overall sentence disclosure. In cases of confusion, redirecting the user to a conversation level is preferable over providing a potentially inaccurate answer.

[x, y: sentence(x), impact(y), outcome(x,y)]

Sentence: the actual input sentence.

Impact: analyzing the impact of each word individually, creating a discourse and generating the outcome of the sentence.

The connection between a pronoun and its antecedent has been extensively studied in linguistics and philosophy. A pronoun is classified as anaphoric, relying on an antecedent expression elsewhere in the sentence or discourse for interpretation, rather than being deictic.

We emphasize anaphoric links through underlining and, considering "beat" as a referential expression, assert that "caught" is also referential in this context, deriving its reference from its antecedent. In cases of ambiguity or multiple scenarios, especially with Large Language Models (LLMs), generating meaningful disclosures or maintaining coherent conversations becomes challenging. SDRT (Segmented Discourse Representation Theory) combined with Attention mechanisms can offer clarity in sentence formation and disclosures by addressing these challenges.

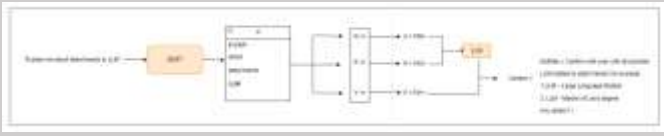


Fig-1 : SDRT based attention mechanisms operate after the disclosure of each word in a tokenized sentence using (Q,K,V) for any given query.

Encoder Architecture:

Models exclusively based on an encoder can be implemented by incorporating SDRT-based attention, segmenting sentences before applying attention. A Discourse Representation Structure (DRS) is a mental construct formed by the hearer as the discourse unfolds, comprising a universe of "discourse referents" representing discussed objects and a set of DRS conditions encapsulating information about these referents. The resulting DRS illustrates the involvement of two individuals—a farmer and a donkey—conveying that the farmer pursued the donkey.

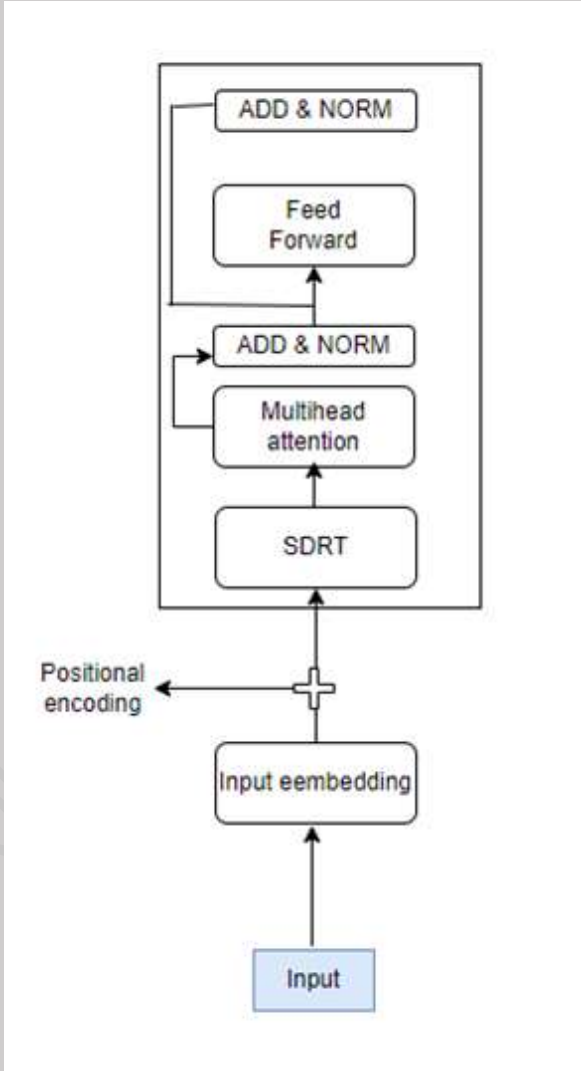


Figure 2

Encoder architecture with SDRT attention based.

$p\theta(x|z) = p\theta(x1, x2, ..., xm|z) = p\theta(x1|z) \cdot p\theta(x2|z) \cdot .... \cdot p\theta(xm|z)$ assume $x1, x2, ..., xm \perp\!\!\!\perp z$

sentence tokenized based on disclosure and attention applied on top of disclosure to check if the machine understood sentence grammar correctly before providing any response.

Architecture

SDRT attention-based Encoder and decoder, will enable grammar-based sentences and questions to be focused on analyzing the relationships between different parts of a sentence and how they contribute to the overall meaning. The key idea behind SDRT is to represent the meaning of a sentence in terms of a set of interdependent semantic structures. These structures are known as semantic dependency trees or graphs, which capture the relationships between words or phrases in a sentence.

The process of SDRT involves the following steps:

Parsing: The sentence is initially parsed into a syntactic structure, such as a constituency tree or dependency tree, which represents the grammatical structure of the sentence.
Semantic Dependency Representation: From the syntactic structure, a semantic dependency representation is derived by identifying the relationships between words or phrases that convey the meaning of the sentence. These relationships are typically labeled with semantic roles, such as agent, patient, location, time, etc. in the given example LLM based on the latest trend indicated with Machine learning based not law based, so this decision needs to be taken by attention and ask the right questions to users.

Reranking: In SDRT, multiple possible semantic dependency structures can be generated from the same syntactic structure. The reranking step involves evaluating and selecting the most likely or preferred semantic dependency structure based on various linguistic and contextual factors. This reranking process helps improve the accuracy and reliability of the semantic representation.

End to end transformer architecture with SDRT attention

Decoder only feature with SDRT

Decoder is one of the key components of the model architecture. It is responsible for generating output sequences based on the encoded input.

In a transformer architecture, the decoder receives encoded representations from the encoder and utilizes self-attention mechanisms along with feed-forward neural networks to generate the output. The decoder typically consists of multiple layers, each containing self-attention and feed-forward sub-layers.

The self-attention mechanism in the decoder allows it to attend to different parts of the input sequence during the decoding process. This attention mechanism helps the decoder focus on relevant information and capture dependencies between different positions in the input.

The output of the decoder is usually a probability distribution over the vocabulary, indicating the likelihood of each word or token in the output sequence.

Before SoftMax calculation done on each token, word, SDRT with disclosure of each word weight will be calculated so that the attention mechanism can act based on weight and make more informed decisions.

The self-attention mechanism in a transformer calculates attention weights for each token by comparing it to all other tokens in the input sequence. This can be represented mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T)/\sqrt{d_k}) \cdot V$$

With SDRT

$$\text{Attention}(Q, K, V, SD) = \text{softmax}((QK^T)/\sqrt{d_k}) \cdot V + SD$$

Here, Q, K, and V are the queries, keys, and values, respectively. They are obtained by linear projections of the input tokens:

$$Q = XW\_Q$$

K = XW_K

V = XW_V

SD = XW_SD

W_Q, W_K, and W_V & XW_SD are learnable weight matrices used for the projections.

After calculating the attention weights, a weighted sum of the values is computed to obtain the attended representation:

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T)/\sqrt{d_k}) \cdot V + SD$$

Finally, the output of the decoder is obtained by passing the attended representation through a feed-forward neural network.

Apart from this more simplified formula for transformers will be the output of SDRT with attention and feed forward neural network will be key components of the model.

Encoder only feature with SDRT

Encoder functions, Let's consider an input sequence $X = [x_1, x_2, ..., x_n]$, where x_i represents the i-th token in the input sequence.

The encoder in a transformer model consists of multiple layers, each containing two sub-layers: a multi-head self-attention mechanism and a position-wise feed-forward neural network.

1. Multi-Head Self-Attention:

The multi-head self-attention mechanism in the encoder allows it to capture dependencies between different positions in the input sequence. It can be represented mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T)/\sqrt{d_k}) \cdot V$$

Here, Q, K, and V are obtained by linear projections of the input tokens:

Q = XW_Q

K = XW_K

V = XW_V

W_Q, W_K, and W_V are learnable weight matrices used for the projections.

With SDRT weight

$$\text{Attention}(Q, K, V, SD) = \text{softmax}((QK^T)/\sqrt{d_k}) \cdot V + SD$$

Where SD – segment disclosure of each token in the given context of the grammar before it passes to the next level.

After calculating the attention weights, a weighted sum of the values is computed to obtain the attended representation.

2. Position-Wise Feed-Forward Neural Network:

After obtaining the attended representation, it is passed through a position-wise feed-forward neural network. This network applies a non-linear transformation to each position independently. It can be represented as:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

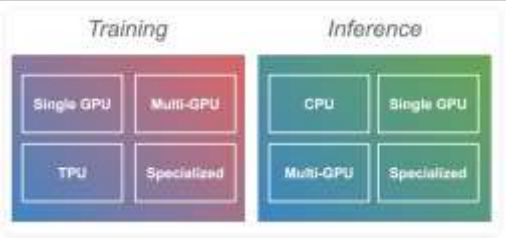
W_1, b_1, W_2, and b_2 are learnable weight matrices and bias terms used in the feed-forward neural network.

These two sub-layers in the encoder are typically followed by layer normalization and residual connections to improve the flow of information.

The encoder can be represented as a composition of these sub-layers across multiple layers. Each layer receives the output of the previous layer as input and passes its output to the next layer.

Performance and Improvements

There will be some performance challenges due to extra calculation of weight, but it can be minimized once tokenization is done. For large language models with multiple sentences will need TPU.



X = Self

SYS = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)

NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA node

PHB = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)

PXB = Connection traversing multiple PCIe bridges (without traversing the PCIe Host Bridge)

PIX = Connection traversing at most a single PCIe bridge

Existing transformers training larger and larger, with SDRT + transformer models and deploying them to production comes with a range of challenges. During training our model can require more GPU memory than is available or be very slow to train and when you deploy it for inference it can be overwhelmed with the throughput that is required in the production environment.

Training and validation

We used a python based hyperparameter- Hyperparameter Search backend, this will help to get fine tuned hyperparameters required for the model to perform at best.

```
pip install optuna/sigopt/wandb/ray[tune]

def SDRT_LLMspace(trial):

   return {

  "learning_rate": trial.suggest_float("learning_rate", 1e-6, 1e-4, log=True),

  "per_device_train_batch_size": trial.suggest_categorical("per_device_train_batch_size", [16, 32, 64, 128]),
```

Conclusion

In this paper we introduced SDRT based attention models for transformers, and it is going to solve most of the biases, providing wrong answers, or misleading answers with the help of proper disclosures and conversational based engines. We also trying to address limitations of attention models and why we need disclosures based attention before providing final output, due to growing complexity of NLP, NLU in the complex environment, transparency will be critical for LLMs, using SDRT based attention will solve most of the problems, but at same time we need lot of data to test and performance also needs to be improved, as next step we are trying to make flash-attention with SDRT to make performance better and fast.

References

[1]Simidjievski, Nikola; Bodnar, Cristian; Tariq, Ifrah; Scherer, Paul; Terre, Helena Andres; Shams, Zohreh; Jamnik, Mateja; Liò, Pietro. "Variational autoencoders for cancer data integration: Design principles and computational practice." Frontiers in Genetics, 10 (2019). ISSN: 1664-8021. doi: 10.3389/fgene.2019.01205.

[2]Asher, Nicholas. Book on SDRT published by Cambridge University Press. https://homepages.inf.ed.ac.uk/alex/sdrt.html.

[3]Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia. "Attention Is All You Need." https://arxiv.org/abs/1706.03762.

[4]Bouchacourt, Diane; Tomioka, Ryota; Nowozin, Sebastian. "Multi-Level Variational Autoencoder: Learning Disentangled Representations From Grouped Observations." Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), April 2018. ISSN: 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11867. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11867.

[5]Thapa, Surendra Bikram; Adhikari, Surabhi. "ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls." Annals of Biomedical Engineering, 2023.